Applying genetics in inflammatory disease drug discovery

Folkersen, Lasse; Biswas, Shameek; Frederiksen, Klaus Stensgaard; Keller, Pernille; Fox, Brian; Fleckner, Jan

Published in:
Drug Discovery Today

Link to article, DOI:
10.1016/j.drudis.2015.05.012

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

feature

Applying genetics in inflammatory disease drug discovery

Lasse Folkersen [1,2], lassefolkersen@gmail.com, Shameek Biswas[3], Klaus Stensgaard Frederiksen[1], Pernille Keller[1], Brian Fox[3] and Jan Fleckner[1]

Recent groundbreaking work in genetics has identified thousands of small-effect genetic variants throughout the genome that are associated with almost all major diseases. These genome-wide association studies (GWAS) are often proposed as a source of future medical breakthroughs. However, with several notable exceptions, the journey from a small-effect genetic variant to a functional drug has proven arduous, and few examples of actual contributions to drug discovery exist. Here, we discuss novel approaches of overcoming this hurdle by using instead public genetics resources as a pragmatic guide alongside existing drug discovery methods. Our aim is to evaluate human genetic confidence as a rationale for drug target selection.

Introduction

The validation of preclinical drug candidates for diseases relies on data from several methods. Knockdown animal models, *ex vivo* studies, *in vitro* cell studies, and *in vivo* tissue samples from patients all contribute to preclinical evaluation of the potential of a drug in the treatment of disease (Table 1). However, a clinical trial is required to generate the necessary evidence in humans. As a result of cost and ethical considerations, only drug candidates with the highest likelihood of effecting disease improvement are tested in clinical trials. Although there is no consensus of what this highest likelihood is, human genetics has been suggested as a fourth type of evidence of preclinical drug targets that can be used to examine causality in humans [1].

Notable examples of the value of human genetics in drug discovery include rare and common proprotein convertase subtilisin/kexin type 9 (PCSK9) variants, for which several drug candidates are already undergoing phase III clinical trials to lower low-density lipoprotein (LDL) levels [2–4]. However, for most genetic-effect variants, there is little understanding of the protein or molecule that mediates the effect. For many GWAS-based discoveries, this deficiency is attributed to most disease-associated single nucleotide polymorphisms (SNPs) being non-coding and intergenic but having regulatory potential [5,6].

The primary approach in overcoming this problem is to combine genetic data with measurements of the molecule *in vivo* in humans; for example, studying the variation in the levels of a drug target candidate, such as high-density lipoprotein (HDL), interleukin (IL)-6, and secretory phospholipase A2-IIa (sPLA2-IIa) [7–9]. One hypothesis opines that the existence of an SNP that causes increased levels of a molecule and also increases the risk of disease is strong evidence of the causality of the molecule in disease. Based on such evidence, IL6 inhibitors are being investigated in clinical trials for the treatment of cardiovascular disease. Similar studies have provided new evidence for the causal and noncausal involvement of LDL and HDL in cardiovascular disease. SNPs associated with LDL levels present a higher risk of cardiovascular disease, whereas SNPs associated with HDL have no effect on disease. This corresponds well with the findings that drugs attempting to target HDL have had little success, whereas LDL modulators (statins) are among the most efficacious drugs in the treatment of cardiovascular disease [7].

The combination of genetic data and drug target levels is termed 'Mendelian randomization' and aids in determining the causality of target molecules in a disease [10]. It is essentially analogous to the clinical trial. Rather than

http://dx.doi.org/10.1016/j.drudis.2015.05.012

TABLE 1

Human genetics compared with other preclinical assessment methods

Method	Strength	Weakness
Animal models	Show causal relations	Nonhuman model
Cellular and *ex vivo* models	Show causal relations in human cells	Does not capture whole-organism complexity
***In vivo* expression**	Observed in humans; whole organism	Does not show causality
Human genetics	Shows causal relations in humans; whole organism	Observational

varying the dose in a clinical trial, this approach uses random genetic variations to alter the presence of a target molecule. Thus, human genetics can be considered Nature's own randomization experiment, albeit with millions of independent tests.

So, why is this approach not being adopted widely in drug discovery? One likely answer is cost. Given the number of samples that is necessary to reach definitive conclusions, such measurements can be as expensive as a phase II clinical trial, rendering it unfeasible for use in applied pretrial drug target research. Another possible issue is that many genetic targets inherently are undruggable per accepted criteria [11].

Here, we discuss a pragmatic approach to this problem: reversing the issue and instead analyzing a candidate pipeline using current genetic and genomic resources, based on mRNA expression instead of protein levels. This method was applied to 14 established drug targets in inflammatory autoimmune disease and 12 undisclosed drug targets developed at Novo Nordisk A/S (http://www.novonordisk.com/). Instead of *de facto* drug discovery, the ultimate purpose of this analysis was to prioritize drug targets, using the choice of indication as a significant parameter. Given the current high attrition rates of drug candidates being evaluated in phase II clinical trials, this method could strengthen the choice of indication for a given drug target and, thus, elevate the number of drug targets passing clinical phase II trials.

Practical implementation scheme

The principal resources in our implementation were data on the link between an SNP and a disease and the correlation between an SNP and target molecule expression. The former were obtained from publically available GWAS and, for the latter, we examined the association with gene expression using expression quantitative trait loci (eQTL) databases, in which genotypes and gene expression in relevant tissues are profiled. These choices enabled us to establish a complete *in silico* analysis pipeline, which is crucial for practical use in a drug company.

In total, seven GWAS and six eQTL databases were used. The *P* values from the GWAS were obtained from dbgap [12]. The studies examined included rheumatoid arthritis (RA) [13], Crohn's disease (CD) and ulcerative colitis (UC) [14], systemic lupus erythematous (SLE) [15,16], psoriasis (PSO) [17], and type 1 diabetes mellitus (T1DM) [18], broadly corresponding to the disease areas that are being focused on for the drug candidates. The eQTL data were derived from studies on individual-level expression and genotype data from relevant tissue and cell types, including intestinal biopsies [19], monocytes and B cells [20,21] Collectively, the data covered 177 795 patients.

The study setup is illustrated in Fig. 1. First, SNPs that affect the expression level of the gene that encodes a drug target or its ligand were identified. Then, the disease-related risk of the identified SNPs was evaluated. The detection of such risk supports the causal involvement of the gene in that disease. Here, 'causality' is presented as cases in which the genetic modulation of the mRNA expression level correlates with the risk of disease, similar to how drug treatments modulate the level or signaling of a target to influence disease severity.

For each drug target ligand, all SNPs within 200 kb of the transcribed regions of the gene were queried for their association with gene expression levels in all available eQTL data sets. This limit was arbitrary but based on earlier publications, because most effects were observed in this range [22,23]. Drug target receptors could also be relevant to investigate; however, in the candidates investigated here, most had no eQTL effects; therefore, ligands were chosen as the primary informant. The association between gene expression and genotype was calculated using a linear additive model. If any SNP was associated with gene expression (eQTL), its disease association in GWAS was also examined (including that of neighboring SNPs in linkage disequilibrium).

GWA *P* values were used directly as downloaded from each GWAS. *P* values for eQTL effects were calculated according to an additive linear model. The false discovery rates (FDR) were calculated by rerunning the algorithm on a random set of genes. This was done 1000 times on sets of genes of equal size to the test set. As shown in Fig. 3d,e a variation of the *P* value threshold changed the percentage of genes that have eQTL-SNPs (step 1, left side of figure) and the percentage of these eQTL-SNPs that have GWAS effects (step 2, right side of figure). Based on this, the cutoff was calculated as $P = 6e{-}4$, because this resulted in 5% of random genes with eQTL effects qualifying as potential drug targets. This calculation served as the basis for all algorithm conclusions and is summarized in the scale in Fig. 2c.

Application and discussion

Figure 2 shows the results of applying the human genetics method to 14 autoimmune drugs that are currently approved and used in the treatment of disease, as well as 12 drug target candidates that are under development. Our genetics method supported many established drug candidates for their primary indications; for example, tumor necrosis factor (TNF) blockers, tocilizumab (IL6 receptor; IL6R), and abatacept (cytotoxic T-lymphocyte-associated protein 4; CTLA4) are recommended for testing in RA, with strong genetic significance. The strength of the evidence was estimated using FDRs that were based on studies of randomly selected genes in the same analysis pipeline, as described in above and in Fig. 3.

Tofacitinib (Janus kinase 3; JAK3) and imiquimod (Toll-like receptor 7; TLR7) did not show genetic evidence according to this method, which is explained by the lack of transcriptional regulation and sex chromosomal location of the target genes, respectively. This is shown in the left column in Fig. 2, labeled 'eQTL', which corresponds to the upper decision fork in Fig. 1.

Conversely, established drugs, such as belimumab, had an eQTL effect but only a weak disease association with its main indication, SLE. Therefore, the results for the 11 drug target candidates were interpreted as merely guidance for the indication. If there was any genetically

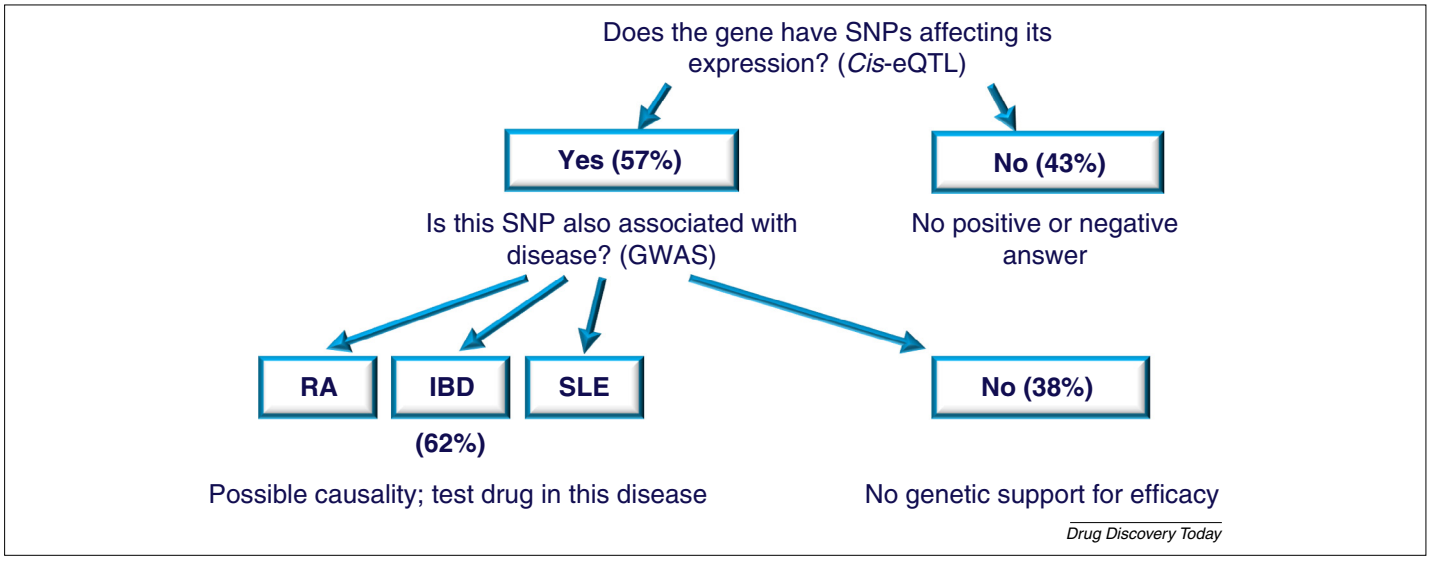

FIGURE 1

Illustration of the human genetics drug assessment scheme. Numbers in parenthesis indicate the percentage of targets sorted at each step.

regulated transcription (eQTL), the corresponding association in GWA data was prioritized. If no eQTL could be identified, the drug target was considered to be intractable with this method.

When applying the method to 12 internal drug discovery targets to evaluate the choice of indication, the recommendations were as follows: Drugs 02 and 06 had a signal for RA (also illustrated in Fig. 2d), whereas Drugs 01, 03, 07, and, to some extent, 05 and 02 had a UC or CD signal. Drug 04 showed a strong signal intensity for SLE as an indication.

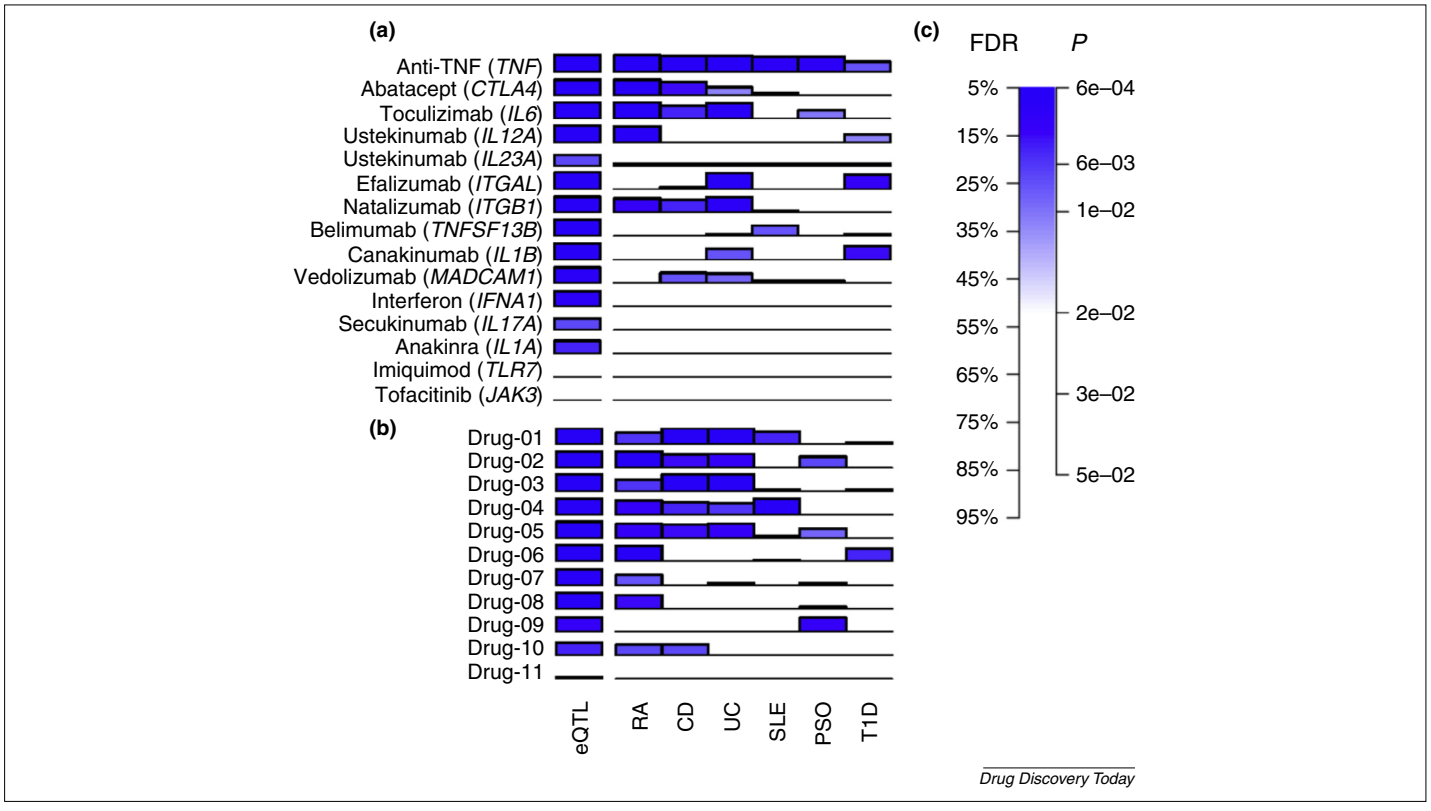

FIGURE 2

Genetic assessment of **(a)** marketed drugs and **(b)** current Novo Nordisk drug targets. The left column shows the best expression quantitative trait loci (eQTL) value for each gene, illustrating cases in which lack of expression or genetic control renders this analysis impossible. The remaining columns show disease association. The principal metric is false discovery rate (FDR), which is shown as the height of the barplot and as a color scale, indicated in **(c)** from 5% to 95%. An FDR of 5% indicates that there is at least one proximal single nucleotide polymorphism (SNP) that affects both disease risk and gene expression of the ligand at a *P* value lower than 5e−4. This value corresponds to the level at which approximately 5% of randomly selected genes show a similar degree of genetic evidence, as further described in Fig. 3.

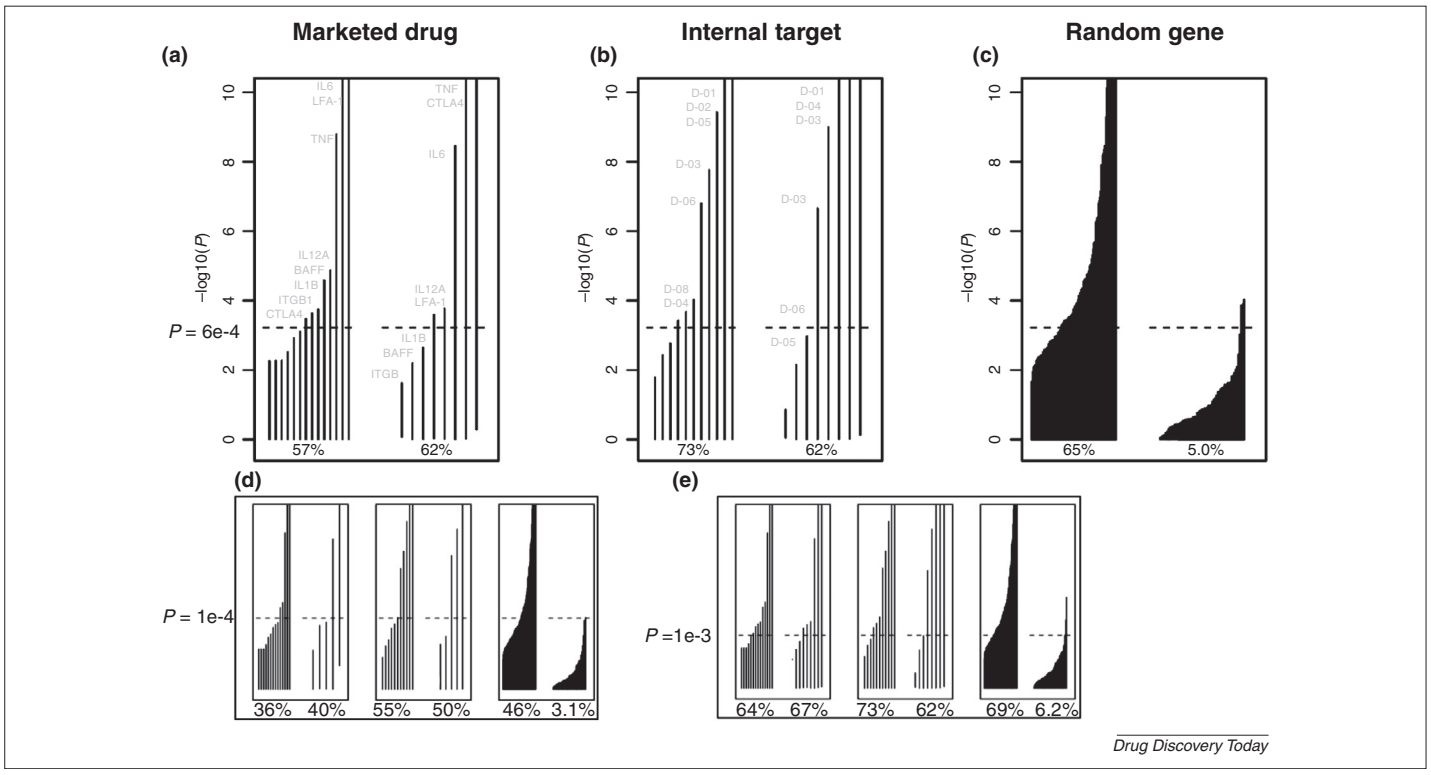

FIGURE 3

Analysis of signal strength from established drug targets and random samplings of genes. In the plots on the left-hand side of each section, each vertical line shows the range of expression quantitative trait loci (eQTL) *P* values for a single gene. These *P* values are calculated from all single nucleotide polymorphisms (SNPs) within 200 kb of the gene in all available eQTL cell and tissue types. The horizontal dotted line indicates the significance threshold of $P < 6e{-}4$. The percentage indicates the fraction of genes that have at least one SNP with a significant eQTL association. In the plots on the right-hand side of each section, each vertical line shows the range of genome-wide association studies (GWAS) *P* values for SNPs surrounding a single gene. However, only SNPs that pass the significance threshold in the left-side eQTL analysis are included. Therefore, a gene is shown only if it had at least one significant eQTL SNP at $P < 6e{-}4$. *P* values from all available autoimmune GWASs are considered. The percentage indicates the fraction of the eQTL-significant genes that also have a GWAS-significant eQTL-SNP. **(a)** Ligands for 14 established autoimmune disease drugs, of which eight have eQTL-significant SNPs. Of these, five have GWAS-significant eQTL-SNPs (as indicated in Fig. 2). **(b)** Ligands for internal targets, with results as described in Fig. 2. **(c)** Results from 1000 random 14-gene grabs from all genes. Of these, 65% have significant eQTL effects, and of these 5.0% have GWAS-associated eQTL SNPs, corresponding to a false discovery rate (FDR) of 5.0%. **(d)** The effect of reducing the *P* value cutoff to 1e−4. This not only reduces the number of targets identified, but also corresponds to an FDR of 3.1%, as seen from a random gene grab. **(e)** Increasing the *P* value cutoff to 1e−3 corresponds to an FDR of 6.2%.

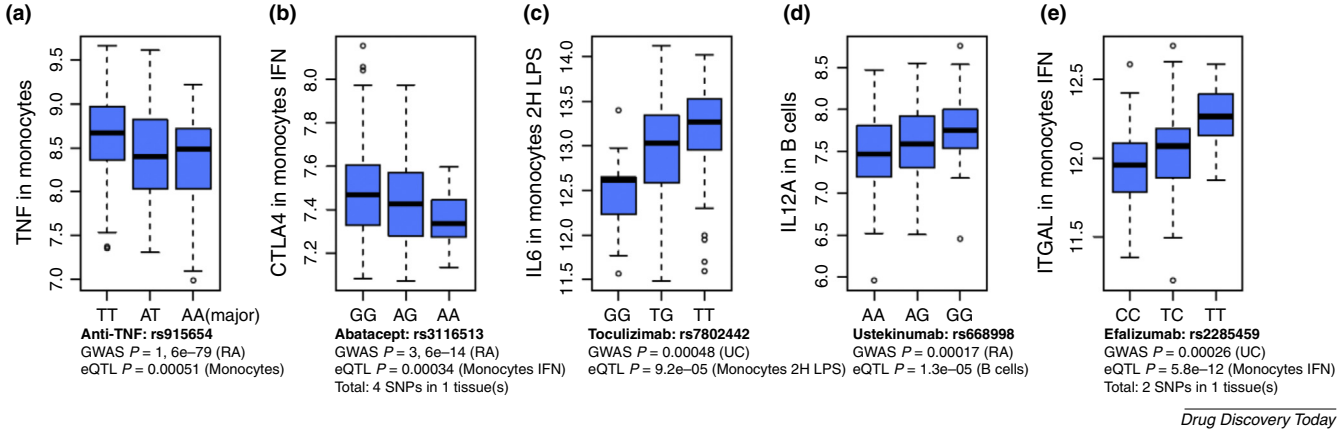

FIGURE 4

Plot of expression levels by single nucleotide polymorphism (SNP) genotype for all marketed drug ligands that pass the $P < 6e{-}4$ threshold both for expression quantitative trait loci (eQTL) and genome-wide association studies (GWAS). Each figure presents one eQTL–GWAS pair. Given that the investigation encompasses multiple SNPs and tissues per drug target, there are two cases where multiple non-linkage disequilibrium (LD) eQTL–GWAS pairs pass the criteria, as indicated for **(b)** and **(e)**. The *Y*-axis indicates mRNA expression level in arbitrary units on a $\log_2$ scale. The *X*-axis indicates genotype, with homozygote risk-allele always shown to the right. Specific *P* values for the eQTL association and the GWAS association are noted below each figure, for the SNP in question or a proxy in high LD, as described in the main text.

Figure 4 highlights the magnitude and direction of effects. It is important to note that, because of compensation mechanisms, it is not inconceivable for eQTL effects to be reversed relative to their effect on protein [24]. Therefore, a complete study of directionality should depend on protein-level measurements. For IL6, we find that a SNP increasing IL6 mRNA expression also increases risk of disease, thereby supporting inhibition of IL6R, consistent with the function of tocilizumab. Similar observations are made for ustekinumab and efalizumab. For abatacept, one would expect that increased levels of CTLA4 would cause increased activation of CD80 and CD86, thereby increasing inflammatory signal and, consequently, disease risk. However, the opposite effect is observed at the mRNA level. Similar remarks can be made for TNF, although being an A/T variant, a strand flip is perhaps a simpler explanation. Therefore, without access to large-scale protein-based Mendelian randomization studies, we recommend to base decisions regarding agonism or antagonism on other study types, preferably including protein-level measurements.

This points to one chief limitation of this method, which is that it only uses mRNA levels, rather than protein levels. Although mRNA and protein levels are often consistent, and this is an assumption that is necessary for performing the presented analysis, there is no guarantee that the protein levels reflect the mRNA levels and, therefore, this is a limitation of the analysis approach. Additionally, the eQTL and GWAS steps are not performed in the same individuals, which is a formal requirement for a study to be termed 'Mendelian randomization'. However, this is the compromise for the requirement of a fully *in silico* process at reasonable expense. Another limitation is that an eQTL SNP sometimes affects the expression of multiple genes; therefore, careful consideration of the SNP and its neighboring SNPs and genes is important. The method is also limited by the current availability of eQTL studies, which might lead to eQTLs that are cell type or condition specific, being missed. The eQTL data are central because few GWAS SNPs are coding, and it is essential to establish a link between gene and SNP, beyond just proximal location. However, as new data become available, the pipeline is updated. We consider all these limitations to be acceptable trade-offs toward using the wealth of large-scale genetic data for drug development.

Likewise, the availability of GWAS data is crucial. However, rather than a limitation, this is a future opportunity: currently the largest GWAS focus on predisposition to disease. As detailed well-powered studies of disease-progression become available, this might shift focus to the use of genetics in drug discovery. However, genetic variants are life long and, therefore, their use might be most relevant for investigations into primary prevention of disease, rather than secondary prevention [25]. Another interesting outlook for the future is the introduction of well-powered rare variant discovery through large-scale sequencing efforts. Although conceptually different, several examples exist where genes with rare coding variants for disease association are also affected by common expression-modulating variants, such as PCSK9 [4]. We believe that future innovations in this field will concern the combination of alternative phenotyping (e.g., disease progression), and novel measurements (rare-variant sequencing). However, here, we want to raise the point that public data sets already contain rich information for use in pragmatic drug-discovery guidance, if not stand-alone discovery.

Concluding remarks

Here, we have proposed and implemented a new method to help select drug target candidates, based on their relevance to specific indications. The relevance of a target is assessed, based on their causal involvement in disease risk. SNPs that cause overexpression of specific genes and are associated with increased disease risk render the gene in question causal in the disease. Thus, altering ligand-induced signaling could decrease the disease risk or severity.

This approach is well founded and constitutes a pragmatic step to enable drug development to benefit from the large investments in genetics that have been made globally in recent years, both by academia and the pharmaceutical industry. This simple concept complements existing drug discovery methods, such as animal models and *in vitro* studies, with disparate advantages and shortcomings. By increasing the confidence of the relevance of a target in a disease, we anticipate that higher-quality drug targets can be selected for further development, which will be valuable if the attrition rate of new drug candidates in clinical proof-of-principle studies remains high.

References

1 Plenge, R.M. *et al.* (2013) Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* 12, 581–594

2 Cohen, J.C. *et al.* (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* 354, 1264–1272

3 Kotowski, I.K. *et al.* (2006) A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am. J. Hum. Genet.* 78, 410–422

4 Chernogubova, E. *et al.* (2012) Common and low-frequency genetic variants in the PCSK9 locus influence circulating PCSK9 levels. *Arterioscler. Thromb. Vasc. Biol.* 32, 1526–1534

5 Encode *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74

6 Farh, K.K. *et al.* (2014) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343

7 Voight, B.F. *et al.* (2012) Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. *Lancet* 380, 572–580

8 Interleukin-6 Receptor Mendelian Randomisation Analysis Consortium (2012) The interleukin-6 receptor as a target for prevention of coronary heart disease: a Mendelian randomisation analysis. *Lancet* 379, 1214–1224

9 Holmes, M.V. *et al.* (2013) Secretory phospholipase A-IIA and cardiovascular disease: a Mendelian randomization study. *J. Am. Coll. Cardiol.* 62, 1966–1976

10 Ebrahim, S. *et al.* (2008) Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Hum. Genet.* 123, 15–33

11 Gashaw, I. *et al.* (2012) What makes a good drug target? *Drug Discov. Today* 17 (Suppl.), S24–S30

12 Mailman *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39, 1181–1186

13 Okada, Y. *et al.* (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376–381

14 Jostins, L. *et al.* (2012) Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119–124

15 Harley, J.B. *et al.* (2008) Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nat. Genet.* 40, 204–210

16 Hom, G. *et al.* (2008) Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. *N. Engl. J. Med.* 358, 900–909

17 Cargill, M. *et al.* (2007) A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes. *Am. J. Hum. Genet.* 80, 273–290

18 Pezzolesi, M.G. *et al.* (2009) Genome-wide association scan for diabetic nephropathy susceptibility genes in type 1 diabetes. *Diabetes* 58, 1403–1410

19 Kabakchiev, B. *et al.* (2013) Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine. *Gastroenterology* 144, 1488–1496

20 Fairfax, B.P. *et al.* (2012) Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* 44, 502–510

21 Fairfax, B.P. *et al.* (2014) Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* 343, 1246949

22 Dimas, A.S. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325, 1246–1250

23 Folkersen, L. *et al.* (2010) Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease. *Circ. Cardiovasc. Genet.* 3, 365–373
24 Maier, T. *et al.* (2009) Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 583, 3966–3973
25 Burgess, S. *et al.* (2012) Use of Mendelian randomisation to assess potential benefit of clinical intervention. *BMJ* 345, e7325

Lasse Folkersen [1,2,*]
Shameek Biswas[3]
Klaus Stensgaard Frederiksen[1]
Pernille Keller[1]
Brian Fox[3]
Jan Fleckner[1]

[1]*Department of PharmacoGenetics, Novo Nordisk, Novo Nordisk Park, Måløv, Denmark*

[2]*Department of Integrative Systems Biology, Center for Biological Sequence Analysis, DTU, Lyngby, Denmark*
[3]*Department of Molecular Immunology, Novo Nordisk, Seattle, WA, USA*

**Corresponding author*: